REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE

14 March 1997

3. REPORT TYPE AND DATES COVERED

Final Report

4. TITLE AND SUBTITLE

Development Of Methods Of Data Preprocessing For Computer Prediction Of New Materials With Predefined Properties

5. FUNDING NUMBERS

F6170896W0296

6. AUTHOR(S)

Dr. Victor Gladun

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

V.M.Glushkov Institute of Cybernetics, National Academy Sciences Ukraine
Prospect Akademika Glushkova, 40
Kiev 252022
Ukraine

8. PERFORMING ORGANIZATION REPORT NUMBER

N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

EOARD
PSC 802 BOX 14
FPO 09499-0200

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

SPC 96-4067

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

A

13. ABSTRACT (Maximum 200 words)

This report results from a contract tasking V.M.Glushkov Institute of Cybernetics, National Academy Sciences Ukraine as follows: The contractor will develop a method for discretization of quantitative data to improve computer prediction of new materials with predefined properties. Develop a software system implementing the method proposed in the above.

14. SUBJECT TERMS

Materials

15. NUMBER OF PAGES

10

16. PRICE CODE

N/A

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19, SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Jerry Sellers

From: Victor P. Gladun
Sent: Friday, March 14, 1997 8:00 PM
To: Jerry Sellers
Subject: final report

Dear Sir,
Thia time I try to send report as a plain text. Unfortunately,
in this format it is impossible to send figures.
Regards, Victor

SPECIAL PROJECT SPC-96-4067
CONTRACT F61708-96-W0296

PREDICTION OF MATERIALS WITH PREDEFINED PROPERTIES.
(final report)

Principal Investigator: Prof.Victor P. Gladun
Co-Principal Investigator: Dr. Neonila D. Vaschenko
Institute of Cybernetics,
Kiev, Ukraine

February, 1997

Contents

1. GOALS OF THE RESEARCH..5
2. PREDICTION OF MATERIALS WITH PREDEFINED PROPERTIES. THE MAIN PRINCIPLES...5
3. THE MAIN PROCESSES OF ATTRIBUTE SET FORMATION.................6
3.1. Man-computer Procedure of Attribute Set Formation..............................6
3.2. Processing of Uncertainties in a Training Set...6
3.3. Discretization of Quantitative Attributes..7
4. REQUIREMENTS TO THE SET OF ATTRIBUTES...................................7
5. REQUIREMENTS TO THE ALGORITHM AND THE PROGRAM OF DISCRETIZATION...

.............................9
6. THE ALGORITHM OF DISCRETIZATION..9
7. COMMENTARY TO THE DISCRETIZATION ALGORITHM..................11
REFERENCES...12
APPENDIX. SUBSYSTEM FOR DISCRETIZATION OF INITIAL DATA (USER'S MANUAL)..13

Abstract

Methods of data preprocessing for computer prediction of new materials are described. The methods result in formation of the attribute set that is used for materials description.

The algorithm of transformation of quantitative attributes into nominal ones (discretization algorithm) is considered in details. It is implemented as a program which is supplied by a USER'S MANUAL.

Subject Terms (Key Words)
Knowledge discovery, diagnostics, prediction, quantitative attributes, discretization, scale transformation.

FOREWORD

This report was prepared by the above identified research team under EOARD Special Contract SPC-96-4067. This is the final report of the contract. We thank Dr(s) Steven R. LeClair, Allen G. Jackson, researchers of the Materials Directorate, Wright Laboratory, and Dr. Nadezda N. Kiselyova, researcher of Moscow Institute of Metallurgy for close collaboration in this work. We also thank Dr.J.Sellers (EOARD) for his support in the administration of the contract.

1. GOALS OF THE RESEARCH

One of the trends involving the design of materials is the use of computational methods for prediction of new materials with predefined properties. Success of the prediction depends to a large extent on quality of an attribute set used for description of chemical compounds.

The research goal is concentrated on the problem of choice and preparation of attribute set to make the results of prediction more reliable. The report contains theoretical foundations concerning formation of attribute set as well as the description of the program supporting this process.

2. PREDICTION OF MATERIALS WITH PREDEFINED PROPERTIES. THE MAIN PRINCIPLES

Prediction of materials with predefined properties is one of the basic processes of new materials design. Prediction is performed on the basis of some general information that characterizes the class of materials having predefined properties. So sometimes it is nesessary to reveal this knowledge about the class by the analysis of known materials having predefined properties as well as materials that are similar to materials with predefined properties but nevertheless do not have these properties. This pr ocess in essence is a learning process. The set of materials that is used for formation of the knowledge is refferred to as a training set.

The knowledge used for prediction must include the most essential combinations of attribute values that usually accompany predefined properties as well as attribute values that can not exist together with predefined properties. It can be described by a logical expression in which essential combinations of attribute values are represented by conjunctions of variables designating attribute values. The logical expression describing a class of materials in essence is its generalized logical model. Formation o f logical models for classes of objects by analysis of training sets is studied

in the framework of such trends as "knowledge discovery", "knowledge mining", "concept learning". The last term defines the process most adequately.
After building generalized logical model for some class of materials prediction of materials of this class is reduced to comparison of its attributive description with the logical expression defining the class. It is performed by calculation of logical expression value after substitution of "1" for variables that are available in the description of the material and "0" for other variables. If the value of logical expression equals "1" the tested material can be created.

3. THE MAIN PROCESSES OF ATTRIBUTE SET FORMATION.

3.1. Man-computer Procedure of Attribute Set Formation

When predicting materials with predefined properties we need set of attributes D={Di} (i=1, 2,...,N) to form attributive descriptions of materials. Each attribute corresponds to set of values Di={dik}, k=1, 2,..., Ki. Attributes can be quantitative, Boolean or nominal. Values of quantitative attributes are numbers. For example, quantitative attributes are temperature, pressure, volume, etc. Values of nominal attributes are some designations. For example, nominal attribute "colour" has values "white", "red", "green" and so on.
The procedure of attribute set formation is based on an understanding of knowledge discovery as an interactive process, each step of which focuses on the refinement of an attribute set, as well as the knowledge that has already been revealed. If the discovered knowledge for some reason is not satisfactory (for example, ineffective for prediction) a researcher changes the attribute set and repeats the process. The form of knowledge representation at the system output should be convenient for the formation of decisions concerning necessary corrections of the attribute set.
In the system CONFOR [1-4] discovered knowledge is represented with a logical expression containing the same designations of attributes that were used in descriptions of the training set. Each conjunction contained in the logical expression is followed by a number indicating how many times it occurs in object descriptions. Similarity in output and input representations simplifies comparison of discovered knowledge with object descriptions and results in formation of ideas concerning improvements in the attribute set.
Changes in the attribute set involve application of the following operations:
Å excluding doubtful attributes,
Å unifying correlating attributes and
Å introduction of new attributes.
Appropriate choice of the attribute set results in formation of concepts of the most simple logical structure.

3.2. Processing of Uncertainties in a Training Set

A training set may include uncertainties of the following types:
1) absence of information about a value of an object attribute;
2) vagueness of information about a value of an object attribute;
3) coincidence of object descriptions belonging to different classes.
Sometimes the degree of vagueness is estimated by the probability of belonging of an attribute value to an object. When applying this approach the main difficulty is connected with the definition of initial probabilities. In practice, initial probabilities are often given only on the basis of voluntary, subjective opinions. Therefore, sometimes it is better to omit doubtful attributes.
Coincidence of object descriptions can be a result of excessive generalization of an attribute. If coincidence of descriptions can not be removed by changing of the attribute set, it is necessary to exclude coincident descriptions from

the training set.

3.3. Discretization of Quantitative Attributes

Logical models of classes as a rule are formed by processing object descriptions involving nominal attributes only. So the problem of transformation of quantitative attributes into nominal ones appears.
This problem is refferred to as a problem of discretization.
Discretization of the scale of quantitative attribute Di means formation of mapping Di -> Bi, where Bi is finite ordered set of intervals of quantitative values.
Bi={bm|(dik (bm): blm(dik<brm}, m=1, 2,...,Mi, where blm and brm are left and right limits of interval bm.
The problem of scale discretization was discussed in scientific literature in connection with processes of classification and formation of linquistic variables for decision making and inference [5-7].

4. REQUIREMENTS TO THE SET OF ATTRIBUTES.

Each classs of objects can be represented by various logical expressions, for example, by different disjunctive normal forms. Naturally, the question appears about the quality of logical models. When such models are used for prediction, the best results correspond, as a rule, to more generalized models that are described by more simple logical expressions. The degree of logical expression simplicity can be measured by the number of its variables.
Quality of logical models formed in the process of learning depends on the set of attributes selected for description of objects and on the method of knowledge formation. To evaluate a set of attributes it is convenient to consider it as N-dimensional space of attributes in which space axes correspond to attributes. In attribute space each object (material) is represented as a point with coordinates that equal to attribute values in its attributive description. Similarity of objects can be evaluated as a distance between points representing these objects. In the space of attributes it is possible to consider distribution areas of objects of different classes as well as surfaces separating these areas.
Let us formulate requirements to the attribute set that influence on the quality of models characterizing classes of objects.

1. Separability of classes.

Sets of objects representing different classes in the training set should be separated in the attribute space. That means that the training set must not include objects of different classes having the same attributive descriptions, i.e. represented in the attribute space by the same point.

2. Compactness of distributions.

Distribution areas of different classes in the attribute space should be compact. The conception of compactness in the attribute space was suggested in [8]. Compactness of distribution of two classes is measured by a number of so called border points. A point representing an object in class A is refferred to as a border point of this class if in the attribute space there are neighbouring points representing objects of other classes. In other words, a point representing an object of class A is a border point if the attributive description of this object can be transformed in a description of some object of another class just by replacing one attribute value by a neighbouring value (for example, by replacing dik by dik-1 or dik+1).
More compact distribution have less border points. Separating surfaces between compact distribution areas are simpler than for noncompact ones. More compact distributions are represented more simple logical models.

3. Simplicity.

The logical model characterizing classes of objects are simpler if a number of attributes and values used for description of objects is not large.
Therefore it is necessary to form an attribute set of the least power, naturally not violating separability of classes. The set of attributes should include only essential attributes characterizing the class as a whole and should not include specific attributes of objects.

5. REQUIREMENTS TO THE ALGORITHM AND THE PROGRAM OF DISCRETIZATION.

1. Discretization must not violate separability of classes in the attribute space.
2. In the attribute space formed as a result of discretization distribution areas of different classes should be as compact as possible.
3. Separability and compactness of classes are provided if the algorithm of discretization reveals the most characteristic intervals in distribution of training set objects on scales of quantitative attributes. The most characteristic intervals are intervals containing objects of the same class as well as intervals with prevalence of objects of some class. So the method of discretization should provide formation of such sort intervals.
4. Under condition of preserving separability and compactness as well as requirement 3 the algorithm of discretization should provide revealing the largest intervals of values. As a result the algorithm will minimize the number of values of nominal attributes.
5. The algorithm of discretization should define limits of characteristic intervals as precisely as possible.
6. The algorithm should allow for a user to generalize the revealed intervals by extension of their limits at the expence of neighbouring empty places. The size of this extension should be defined by a user on the basis of his understanding of the problem for which the discretization is necessary. If the discretization is used for further knowledge discovery and material prediction the degree of interval generalization influences strongly results of the prediction.
7. The program of discretization should include tools for visualization of the process to make it understandable for a user.

6. THE ALGORITHM OF DISCRETIZATION.

Discretization is performed on scales of quantitative attributes by analysis of distributions of training set objects belonging to different classes. Every object of the training set is marked on the scale with its attribute value. In the process of discretization intervals of the following types can be formed:

1) empty, not containing attribute values of objects belonging to the training set;
2) homogeneous, containing marked attribute values of objects of the same class;
3) even, in which there are marked attribute values of different classes and the difference between numbers of marked values of the classes does not exceed the given threshold;
4) uneven, in which there are marked attribute values of different classes and the difference between the number of marked values of some class and the number of marked values of any other one exceeds the given threshold.

The discretization algorithm for attribute Di (i=1, 2,...,N) consists in consecutive preformance of the following operations:

1. Formation of the scale of attribute values describing objects of the training set.

1.1. Definition of the scale limits. The operation involves finding the highest (dimax) and the lowest (dimin) values of the attribute among all objects of the training set.

1.2. Definition of the size d of the initial interval

d = (dimax - dimin)/L, where L is the number of initial intervals.

The algorithm makes it possible choosing L by a user.

1.3. Numbering initial intervals and definition of their limits.

Left and right limits of initial intervals are calculated in such a way:

b1l = dimin; b1r = dimin + d.

bml = bm-1r; bmr = bml+ d; m=2,...,L.

2. Formation of distributions of training set objects.
In each of initial intervals the attribute values belonging to objects of the training set are marked.
Attribute value dik (i=1, 2,...,N, k=1, 2,...Ki) belongs to interval bm (Bi (m=1, 2,...,(Mi -1)) if bml (dik < bmr. For the last interval
bMil
(dik(bMir.

3. Unification of neighbouring empty intervals.
The scale of the attribute is examined from the left limit to the right one. Neighbouring empty intervals are united.

4. Generalization of the distribution.
Nonempty parts of the scale are extended at the expence of joining several empty initial intervals from both sides. The number of initial intervals that should be joined from each side are defined by a user. The operation results in
disappearance of short empty intervals.

5. Unification of neighbouring homogeneous intervals of the same class.
As a result the largest homogeneous intervals are formed.

6. Unification of neighbouring uneven intervals.

7. Unification of neighbouring even intervals.

7. COMMENTARY TO THE DISCRETIZATION ALGORITHM.

The algorithm is aimed at providing separability and compactness of classes in the attribute space. It is achieved by revealing intervals that are the most characteristic for classes, namely, intervals containing attribute values of one class only (homogeneous intervals) and intervals with predominance of attribute values of one class (uneven intervals).
If after discretization the training set includes objects of different classes
having the same descriptions the program CONFOR informs a user about the fact. In this case separability of classes is achieved by correction of intervals by a user or by exclusion of indistinguishable objects of different classes from the training set.
The algorithm is aimed at selection of intervals of the largest size, i.e. at reducing of the total number of nominal attributes being formed. The algorithm reveals intervals that reflect peculiarities of distributions of training set objects and gives a user possibility to generalize revealed intervals on the basis of his understanding of the problem for which the discretization is used.
In contrast to known algorithms of discretization [5-7] the above described algorithm provides completely automatic discretization reflecting peculiarities
of mutual distributions of different classes.

REFERENCES

1. Development of Computer Methods for Prediction of New Materials Having Predefined Properties. EOARD Special Contract SPC-95-4016 (final report), August, 1995.
2. Gladun V.P., Vaschenko N.D. Local-Statistic Methods of Knowledge Discovery. Kibernetika i sistemny analiz, N2,1995 (in Russian, transl. into English).
3. N.N.Kiselyova. "Prediction of Inorganic Compounds: Experiences and Perspectives". MRS BULLETIN, 1993, N2.
4. N.N. Kiselyova. "Information-predicting Systems for the Design of New Materials". Journal of Alloys and Compounds, 197 (1993).
5. Mirkin B.G. Analyses of Qualitative Attributes. Moscow: Statistika, 1976, (in Russian).
6. Markin O.Yu. Discretization on Qantitative Attributes when Preprocessing Information for Generalization. Upravljajuschie sistemi i mashini, N2, 1988 (in
Russian).
7. Lbov G.S. Methods of Processing Polytypic Experimental Data. Novosibirsk: Nauka, 1981 (in Russian).
8. Arkadjev A.G., Braverman A.L. Computer Learning to Classify Objects. Moscow:
Nauka, 1971.

APPENDIX

CONFOR-2

TOOLS FOR KNOWLEDGE DISCOVERY, DIAGNOSTICS AND PREDICTION

Subsystem for Discretization of Initial Data (User's Manual)

1.
Purpose...14
2. Description of Initial
Data..1
14
3.
Main
menu..14
3.1.
Discretization...14
3.1.1. Scale Transformation (for the learning
set).......................................15
3.1.2.
Individual
discretization...16
3.1.3. Scale Transformation (for the recognition or examination
set)..16
3.2.
Concept
Formation...17

1. Purpose.

Initial data for the CONFOR-2 system are descriptions of objects represented as sets of attribute values. Attribute values can be given in quantitative, Boolean or nominal scales. Before using the initial data for knowledge discovery, diagnostics and prediction quantitative attributes should be transformed into nominal ones.
Subsystem for Discretization of Initial Data is a tool for support of preprocessing initial data given in quantitative scales.

2. Description of Initial Data.

Before starting the Confor-2 system you should prepare your initial data as dbase-files and place them into directory with the CONFOR-2 system.
Dbase-files should not include needless information that will not be used in object descriptions. The first field should be an object name, the second field should be a class name, next fields should be attributes. In recognition set the second field (class name) should be filled with "?".
When preparing dbase-files define all fields as symbol strings.
Use 2 symbols for file names:

learning set - n#.dbf,
recognition set - r#.dbf,
examination set - e#.dbf, where # - task name.

Task name is one symbol (letter or figure).

3. Main menu.

The main menu of the CONFOR-2 system consists of 3 buttons:

- Discretization;
- Concept Formation;

- Exit.

The Discretization button opens the window with menu of tasks and makes it possible to call tools for discretization of quantitative attributes and transformation of sets of object descriptions in the form suitable for processes of knowledge discovery, recognition or examination.

The Concept Formation button makes it possible to call tools for concept formation, recognition or examination.

The Exit button enables to quit the system.

3.1. Discretization.

Choose the task of interest. Choose the mode you will work in (learning, recognition or examination) when solving the choosen task.

If nesessary you can delete a task in this window (Delete Task button). Be careful: all dbase-files (n-, r- and e-type) for the choosen task are deleted here.

In the case when you have not discretized the choosen task before, the Scale Transform. and Exit buttons are accessible only.

For the learning set the Scale Transform. button starts the process of formation of intervals for quantitative attributes and, after this, the process of substitution of interval names for attribute values in object descriptions.

For the recognition or examination set the Scale Transform. button starts the process of substitution of interval names (formed for the learning set) for attribute values in object descriptions.

If the choosen task has been discretized before you can see the discretized set (Set View button) or the formed intervals (Intervals view button).

Designations "[", "]" are used for intervals including the limit value and "(", ")" are used for intervals that do not include the limit value.

3.1.1. Scale Transformation (for the learning set).

The Scale Transform. button opens the window with initial object descriptions and menu (Scale Transformation..., Attributes, Exit).

If the learning mode was chosen at the previous step the Scale transformation item of the menu starts the process of interval formation.

The first stage of the process of interval formation is attribute selection.

Use your mouse and Select All button to mark attributes whose values should be discretized.

Formation of intervals is performed on scales of quantitative attributes divided into equal initial intervals. Attribute values of objects belonging to the training set are marked.

In the process of interval formation intervals of the following types can be formed:

1) empty, not containing attribute values of objects belonging to the training set;

2) homogeneous, containing marked attribute values of objects of one class only;

3) even, in which there are marked attribute values of different classes and the difference between numbers of marked values of the classes does not exceed the given threshold;

4) uneven, in which there are marked attribute values of different classes and the difference between the number of marked values of some class and the number of marked values of any other one exceeds the given threshold.

You can use the thresholds by default or set them to the proper values.

"Threshold 1" makes it possible to form uneven intervals with predominance of objects of some class. For uneven intervals Threshold 1 defines a minimal permissible difference between number of objects of predominant class and numbers of objects of another class. By default Threshold 1 equals 2.

"Threshold 2" is used to set the degree of generalization of intervals being formed. It defines the permissible extension of interval limits at the expence of neighbouring empty places. Increasing the threshold favours to extension of formed intervals to the right and to the left. The extension is measured by the number of initial intervals. By default Threshold 2 equals 2. It is a very important adjustment because it influences strongly results of discretized scale application.

"Number of Initial Intervals" defines the initial number of intervals the
algorithm starts from. Increasing this number results in increasing the accuracy of interval formation and decreasing the degree of generalization. By default it equals 100.

Click OK to start the process.
With the process completed, you can see a set of windows with pictures demonstrating transformed scales for each discretized attribute and the objects of the learning set that are described in terms of the interval names.
Use the right top window (Attributes) to see the attribute of interest.
Double clicking the attribute name activizes its window.
To activize the window Attributes use menu on the top of the learning set (item Attributes).
To activize the main window with the learning set press Alt.

3.1.2. Individual discretization.

Activizing the window with the attribute of interest you can influence the
process of discretization of individual attribute by changing intervals with the help of the thresholds. The discretization of an individual attribute can be performed in two ways: automatically or step by step.
For automatic discretization set the thresholds to proper values, and click the
Auto button.
For investigation of the discretization process step by step click the Step button until the message about finishing the process will appear.
The last result will be saved.

3.1.3. Scale Transformation (for the recognition or examination set).

The Scale Transform. button opens the window with initial object descriptions and menu (Scale Transformation, Exit).
Select Scale Transformation to start the process.
With the process completed, you can see the window with the objects of the recognition or examination set described in terms of the intervals that were formed on the basis of the learning set.

3.2. Concept Formation.

Use the Concept Formation button to call tools for knowledge discovery, recognition or examination (see User's manual for the CONFOR system).
You will see in the task menu the names of tasks that you worked with at the stage of discretization.
Be careful: if you want to change intervals for a task and to run Concept Formation once more you should change the task name.

------ Message Header Follows ------
Received: from whale.gu.kiev.ua by smtp-gw.ehis.navy.mil
  (PostalUnion/SMTP(tm) v2.1.9c for Windows NT(tm))

```
id AA-1997Mar14.200052.1027.139381; Fri, 14 Mar 1997 20:00:54 GMT
Received: from creator.gu.kiev.ua (creator.gu.kiev.ua [194.93.190.3]) by
whale.gu.kiev.ua (8.8.5/8.7.3) with ESMTP id WAA156290 for
<jsellers@eoard.af.mil>; Fri, 14 Mar 1997 22:00:51 +0200
Received: from aduis.UUCP (uuoilinf@localhost) by creator.gu.kiev.ua  with
UUCP
id VAA22054 for jsellers@eoard.af.mil; Fri, 14 Mar 1997 21:54:59 +0200 (EET)
X-Authentication-Warning: creator.gu.kiev.ua: uuoilinf set sender to
aduis!aduis.kiev.ua!glad using -f
Received: by aduis.kiev.ua (dMail for DOS v1.23, 15Jun94);
          Fri, 14 Mar 1997 21:32:20 +0200
To: jsellers@eoard.af.mil
Message-Id: <AB4VQApGh8@aduis.kiev.ua>
Organization: ADUIS
Date: Fri, 14 Mar 1997 21:32:20 +0200 (UKR)
From: "Victor P. Gladun" <glad@aduis.kiev.ua>
X-Mailer: dMail [Demos Mail for DOS v1.23]
Subject: final report
Lines: 287
```